Seminar: Practical Machine Learning

Variational Drop Out

Ashuha Arseniy

Yandex, Moscow Institute of Physics and Technology

ars.ashuha@gmail.com

August 31, 2016

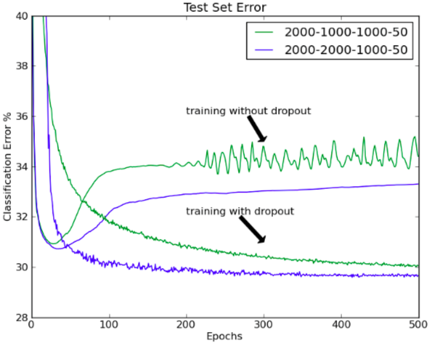
Kingma, et al. "Variational Dropout and the Local Re-parameterization Trick." NIPS'15

Binary Drop Out

Dropout is a technique for regularization of neural network

$$B = (A \circ \xi)\theta$$

A – inputs, $\xi \sim Bernulli(1-p)$, θ – weights, B – outputs



Hinton at el.: Improving neural networks by preventing co-adaptation

Gaussian Drop Out

$$B = (A \circ \xi)\theta, \xi \sim Bernulli(1-p)$$

- During dropout testing we need to scale the weights on

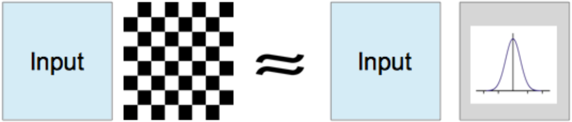
$$\theta_{test} = 1/(1-p)\theta_{train}$$

- The same can be achieved by scale activation on $1/(1-p)$ at training

$$B = (A \circ \xi)\theta \qquad \rightarrow \qquad B = (A \circ \xi/(1-p))\theta$$

$$\mathbb{E}[\xi_{ij}/(1-p)] = 1, Var[\xi_{ij}/(1-p)] = p/(1-p)$$

- Using a distribution with the same mean and variance, works as well



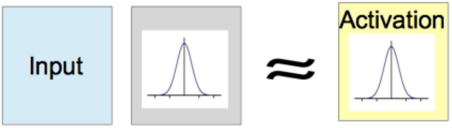
Fast Gaussian Drop Out

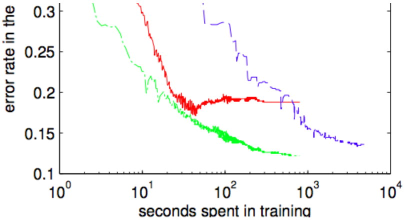
Gaussian dropout: $$B = (A \circ \xi)\theta$$

A – input matrix, $\xi \sim N(1, \alpha)$, θ – weight matrix, B – output matrix

- It means that activation also distributed Normal

$$q(b_{mj}|A, \theta, \alpha) = N(\gamma_{mj}, \delta_{mj}) \quad \gamma_{mj} = \sum_i a_{mi}\theta_{ij}, \delta_{mj} = \alpha \sum_i a_{mi}^2 \theta_{ij}^2$$





- It's equivalent Normal distribution on weight $q(w_{ij}) = N(\theta_{ij}, \alpha\theta_{ij}^2)$

$$b_{ij} = \sum_k a_{ik}(1 + \sqrt{alpha} \cdot \epsilon)\theta_{kj}, \quad \epsilon \sim N(0, 1), N(\mu, \sigma^2) = \mu + \sigma\epsilon$$

Wang at el.: Fast dropout training

Mind Summary

- We transform α from hyper-parameter to parameter
- Also transform model to probabilistic
- We want to use stochastic variational toolbox
- Our goal is to see that dropout is special case of Bayesian Inference

Variational Lower Bound

We want to determinate parameters $\phi = (\theta, \alpha)$ of distribution on $Z = w_{ij}$

Let's use Bayesian toolbox:

1. Defined $p(X, Z)$, X – observed variables, Z – hidden variables
2. We want to find $p(Z|X) = P(X, Z)/P(X)$, $P(X)$ usually intractable
3. Therefor we will try to approximate $p(Z|X) \approx q_\phi(Z)$
4. Derive Variational Lower Bound

$$log\ P(X) = \int q_\phi(Z) log\ P(X) dZ = \int log\ \frac{p(X,Z)q_\phi(Z)}{p(Z|X)q_\phi(Z)}\ q_\phi(Z) dZ =$$
$$\int log\ \frac{p(X,Z)}{q_\phi(Z)}\ q_\phi(Z) dZ + \int log\ \frac{q_\phi(Z)}{p(Z|X)}\ q_\phi(Z) dZ =$$
$$= \mathcal{L}(q_\phi(Z), p(X, Z)) + D_{KL}(q_\phi(Z), p(Z|X))$$

5. $$\mathcal{L} = \mathbb{E}_{q_\phi(Z)} log\ p(X|Z) - D_{KL}(q_\phi(Z), p_{prior}(Z))$$

likelihood expectation regularizer

Stochastic Gradient Variational Lower Bound

$$\mathcal{L} = \mathbb{E}_{q_\phi(Z)} log\ \ p(X|Z) - D_{KL}(q_\phi(Z), p_{prior}(Z))$$

1. We cant take gradient by parameters by Naive way
2. The problem is in estimate this derivative

$$\frac{\partial}{\partial \phi}\mathbb{E}_{q_\phi(Z)} log\ \ p(X|Z) \neq \mathbb{E}_{q_\phi(Z)}\frac{\partial}{\partial \phi} log\ \ p(X|Z)$$

3. Lets use re-parametrization trick:

$$\frac{\partial}{\partial \phi}\mathbb{E}_{q_\phi(Z)} log\ \ p(X|Z) = \mathbb{E}_{N(\epsilon|0,1)}\frac{\partial}{\partial \phi} log\ \ p(X|Z = f(\epsilon, \phi))$$

4. Example:

$$f(\epsilon, (\theta, \alpha)) = \theta + \sqrt{\alpha}\theta\epsilon$$

Kingma, et al. "Auto-Encoding Variational Bayes"

Variational Drop Out

1. During dropout training we optimize Expectation on MLE

$$\mathbb{E}_{q_{\alpha,\theta}} log\ p(X|\alpha,\theta) \to \max_{\theta}$$

2. During dropout training we optimize VLB with fixed α, $\alpha = constant$

$$\mathbb{E}_{q_{\alpha,\theta}} log\ p(X|\alpha,\theta) + D_{KL}(\alpha) \to \max_{\theta,\alpha}$$

$$\mathcal{L} = \mathbb{E}_{q_\phi(Z)} log\ p(X|Z) - D_{KL}(q_\phi(Z), p_{prior}(Z))$$

3. There exist the $p(log(|w_{ij}|)) \propto c$ prior satisfy this conditions
4. Prior interpretation – number of significant digits

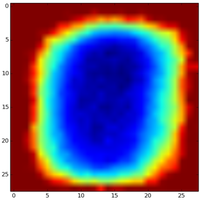
We can train personal alpha for: weight, features, layer

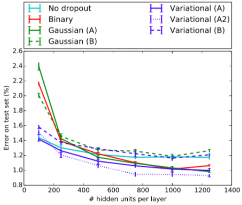
- Uncorrelated – separate alpha per weight
- Correlated – separate alpha per weight correspond to same features

$$q(Z) = q(w_{ij}) = N(\theta_{ij}, \alpha_{ij}\theta_{ij}^2)\ or\ N(\theta_{ij}, \alpha_i\theta_{ij}^2)$$

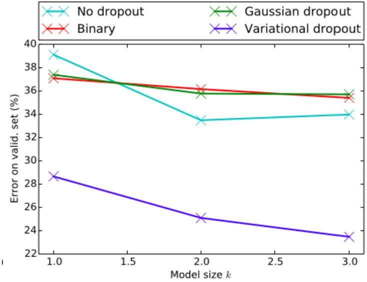
Kingma, et al. "Variational Dropout and the Local Reparameterization Trick"

1. MNIST





2. CIFAR



Summary

- Bernoulli dropout can be transform to Gaussian with same E, Var
- Gaussian input noise equal Gaussian weight noise
- Reinterpretation weight Gaussian noise as posterior distribution
- Show that dropout training is special case of Bayesian Inference
- Offered efficient low variance gradient estimator